Домашнее задание №2.

Домашнее задание сдается в электронном виде в SmartLMS.

Срок сдачи: 24 ноября 2024 года, 23:59.

На титульном листе обязательно указать:

Фамилию И.О., номер варианта.

Домашнее задание основано на результатах опроса населения РМЭЗ НИУ ВШЭ в 2020, 2022 и 2023 годах (https://www.hse.ru/rlms/spss). В файле $Homework\_2\_data.csv$ (файл CSV) содержатся следующие переменные:

- wage — заработная плата, полученная за последние 30 дней по основному месту работы после удержания налогов в рублях;
- educ — уровень образования, категориальная переменная (0 для индивидов, учившихся в школе):
  1. ПТУ, техническое училище
  2. институт, университет, академия
- female = 1, если респондент – женщина, = 0 для мужчин;
- age — возраст в годах;
- is_children = 1, если у респондента есть хотя бы 1 ребенок, = 0 иначе;
- work_hours — количество часов, которое продолжается рабочий день;
- foreign_language = 1, если респондент знает ли иностранный язык, = 0 иначе;
- internet = 1, если респонденту приходилось в течение последних 12 месяцев пользоваться Интернетом, = 0 иначе;
- alcohol = 1, если респондент употребляет алкогольные напитки (хотя бы изредка), = 0 иначе;
- health = 1, если респондент испытывал проблемы со здоровьем за последний месяц, = 0 иначе;
- weight — вес респондента в кг;
- height — рост респондента в см;
- smoke = 1, если респондент курит, = 0 иначе;
- industry — отрасль занятости:

1. ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ
2. ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ
3. ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС
4. НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ
5. ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ
6. СТРОИТЕЛЬСТВО
7. ТРАНСПОРТ, СВЯЗЬ
8. СЕЛЬСКОЕ ХОЗЯЙСТВО
9. ОРГАНЫ УПРАВЛЕНИЯ
10. ОБРАЗОВАНИЕ
11. НАУКА, КУЛЬТУРА
12. ЗДРАВООХРАНЕНИЕ
13. АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ
14. ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ
15. ФИНАНСЫ
16. ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ
17. ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО
18. ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ
19. СОЦИАЛЬНОЕ ОБСЛУЖИВАНИЕ
20. ЮРИСПРУДЕНЦИЯ
21. ЦЕРКОВЬ
22. ХИМИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ
23. ДЕРЕВООБРАБАТЫВАЮЩАЯ ПРОМЫШЛЕННОСТЬ, ЛЕС
24. СПОРТ, ТУРИЗМ, РАЗВЛЕЧЕНИЯ
25. УСЛУГИ НАСЕЛЕНИЮ
26. IT, ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
27. ЭКОЛОГИЯ, ЗАЩИТА ОКРУЖАЮЩЕЙ СРЕДЫ
28. ОРГАНИЗАЦИЯ ОБЩЕСТВЕННОГО ПИТАНИЯ
29. СМИ, ИЗДАТЕЛЬСТВО, ПЕЧАТЬ, ТЕЛЕКОММУНИК
30. РЕКЛАМА, МАРКЕТИНГ
31. ОБЩЕСТВЕННЫЕ ОРГАНИЗАЦИИ, СОВЕТ ВЕТЕРАН

- regions — регион проживания респондента;
- year — год проведения опроса;

Используя выбранные для Вас данные (см. год, отрасль и регион в ведомости), выполните приведенные ниже упражнения.

Для отбора варианта — исполните следующий код, вставив вместо пропусков строки с назначенными Вам годом, отраслью или регионом.

```
import pandas as pd
df_hw = pd.read_csv('Homework_2_data.csv')
year = # your text
industry = # your text
region = # your text
if industry == '       ':
    my_data = (df_hw[(df_hw.year == year) &
                     (df_hw.region == region)])
elif region == '       ':
    my_data = (df_hw[(df_hw.year == year) &
                     (df_hw.industry == industry)])
my_data
```

В случае сдачи работы не своего варианта (расчеты не будут соответствовать выборке из варианта), оценка за работу составит 0 баллов.

Внимание! Для проверки гипотез требуется сформулировать нулевую и альтернативную гипотезы, а также указать расчетное (формулу расчета с подстановкой данных, соответствующих Вашей выборке) и критическое значение статистики, а также рассчитать p-value. Расчеты можно делать в Python или иной программе. Результаты всех тестов должны быть проинтерпретированы (отвергается или не отвергается гипотеза, на каком уровне значимости и что это значит). На выходе работа должна быть представлена в виде текста (комментариев), результатов расчетов, графиков и таблиц. Задания в тексте обязательно должны быть пронумерованы, согласно пунктам ниже. Каждый пункт оценивается в 2 балла.

1. Опишите Вашу выборку. Что является в Вашем случае генеральной совокупностью? Как можно проверить репрезентативность Вашей выборки? (описать словами, не проверять).
2. Рассчитайте описательные статистики (минимум, максимум, среднее значение, стандартное отклонение, размах) для всех переменных в Вашей выборке кроме отрасли, региона и года.

3. Оцените квартили (25%, 50%, 75%) распределения для количественных переменных в выборке. Определите межквартильный размах.
4. Сравните среднее значение, медиану и моду для количественных переменных в выборке. Что можно сказать об их соотношении?
5. Постройте box–plot для всех количественных переменных. Есть ли выбросы?
6. Постройте гистограммы распределения для количественных переменных в выборке. Что можно сказать о скошенности (асимметрии) и островершинности их распределений? Рассчитайте соответствующие показатели (Skewness и Kurtosis) и сделайте выводы.
7. Как распределены респонденты в Вашей выборке по уровню образования? Постройте гистограмму.
8. Постройте корреляционную таблицу для всех переменных в Вашей выборке кроме отрасли, региона и года. Проинтерпретируйте результаты.
9. Предположите зависимость заработной платы от каких-либо переменных в файле. Постройте графики, которые позволяют продемонстрировать эту зависимость.
10. Оцените линейную модель, которая объясняет заработную плату ($wage$) возрастом (age), наличием высшего образования ($high$), полом ($female$), наличием детей ($is\_children$), курением ($smoke$) и константой. Проинтерпретируйте полученные результаты. Все ли коэффициенты оказались значимы? Выпишите уравнение оцененной модели.
11. Выполните тест на адекватность этой модели и сделайте выводы.
12. Сформулируйте и протестируйте гипотезу для одного из коэффициентов модели. Дайте содержательную и количественную интерпретацию полученных результатов.
13. Сформулируйте и протестируйте гипотезу о нескольких коэффициентах модели. Дайте содержательную и количественную интерпретацию полученных результатов.
14. Постройте график «остатки–прогнозы». Сделайте вывод.
15. Оцените модель из п. 10, оставив в ней только значимые коэффициенты. Выпишите уравнение оцененной модели. Сравните результаты с моделью из п. 10. Какие критерии для сравнения моделей здесь стоит использовать?
16. Протестируйте наличие выбросов в модели с помощью известных Вам методов. Если они есть, то как их учесть в модели? Проведите коррекцию.
17. Постройте прогноз заработной платы для одного индивида с заданными вами характеристиками для него. Постройте 95% доверительный интервал для прогнозного значения.